ISSN 1229-1889(Print)
ISSN 2287-9005(Online)
J Korean Soc Qual Manag Vol. 47, No.2: 271-281, June 2019
https://doi.org/10.7469/JKSQM.2019.47.2.271

연관성 기반 비유사성을 활용한 범주형 자료 군집분석

이창기 · 정욱†

동국대학교 경영대학

Categorical Data Clustering Analysis Using Association-based Dissimilarity

Changki Lee · Uk Jung†

College of Business Administration, Dongguk University

ABSTRACT

Purpose: The purpose of this study is to suggest a more efficient distance measure taking into account the relationship between categorical variables for categorical data cluster analysis.

Methods: In this study, the association-based dissimilarity was employed to calculate the distance between two categorical data observations and the distance obtained from the association-based dissimilarity was applied to the PAM cluster algorithms to verify its effectiveness. The strength of association between two different categorical variables can be calculated using a mixture of dissimilarities between the conditional probability distributions of other categorical variables, given these two categorical values. In particular, this method is suitable for datasets whose categorical variables are highly correlated.

Results: The simulation results using several real life data showed that the proposed distance which considered relationships among the categorical variables generally yielded better clustering performance than the Hamming distance. In addition, as the number of correlated variables was increasing, the difference in the performance of the two clustering methods based on different distance measures became statistically more significant.

Conclusion: This study revealed that the adoption of the relationship between categorical variables using our proposed method positively affected the results of cluster analysis.

Key Words: Association-based Dissimilarity, Distance Metric, Unsupervised Learning, Categorical Data, Clustering

● Received 10 March 2019, 1st revised 24 March, accepted 25 March 2019
† Corresponding Author(ukjung@dongguk.edu)

※ 이 연구는 2018년도 한국연구재단의 국제협력사업의 지원(2017K2A9A2A06016127)과 2019학년도 동국대학교 논문게재장려금 지원으로 이루어졌음.

1. 서 론

4차 산업혁명의 시대가 시작됨에 따라 기업들은 정보를 수집 및 분석하고 이를 품질경영에 활용하는 것을 기업 경쟁력 향상의 중요 요인으로 여기고 있다. 그러나 IT 기술의 발전으로 수집되는 데이터의 규모가 매우 크고 복잡해짐에 따라 전통적인 통계적 분석방법 이외에 다른 분석방법의 필요성이 증대되었다. 이에 따라 대용량의 복잡한 데이터로부터 유용한 정보를 추출하고 이로부터 의미 있는 규칙이나 패턴을 찾는 데이터마이닝(Datamining)과 기계학습(Machine Learning) 분석방법이 대안으로 제시되었다(Seo and Yun 2017; Lim et al. 2016)

품질경영 영역에 데이터마이닝과 기계학습을 활용하는 과정에서 두 관측치(observation) 사이에 비유사도(Dissimilarity) 혹은 거리(Distance)를 측정하는 행위는 데이터로부터 의미 있는 규칙이나 패턴을 찾는 과정에서 매우 중요한 역할을 한다(Jia et al. 2016). 일반적으로, 연속형 변수(Continuos variables or Quantitative variables)로 측정된 관측치의 거리는 유클리디언 거리(Euclidean distance), 민코프스키 거리(Minkowski distance), 코사인 거리(Cosine distance)와 같이 많은 연구자에 의해 오랜 시절에 걸쳐 다양한 방법이 제시됐다 (Esposito et al. 2000; Cha 2007; Jia et al. 2016). 이에 비해 범주형 변수(Categorical variables or Qualitative variables)에 관해서는 상대적으로 더디게 연구가 진행되어 왔다. 과거 제조업 중심의 산업에서 얻어지는 관측치는 대부분 연속형 변수가 주를 이루어 왔지만, 최근 서비스 산업의 비중이 증가함에 따라서 과거와 달리 연속형 변수로 측정된 관측치가 아닌 범주형 변수로 측정된 관측치를 수집하는 경우가 증가하고 있다. 예를 들면, 의료업에서 고객의 정보를 수집하는 경우 성별, 기혼 여부, 가족 병력 유무 같은 경우가 이에 해당한다. 이 경우 앞서 말한 유클리디언 거리와 같은 연속형 변수의 거리를 측정하기 위한 측정 방법은 적용할 수 없다.

범주형 변수로 측정된 관측치의 거리를 정의하기 위해 제안된 가장 간단한 방법은 해밍 거리(Hamming 1950)이다. 해밍 거리는 단순하게 변수 값이 같으면 0 다르면 1로 구분하여 두 변수 값의 비유사도를 계산하고 이 비유사도의 합을 두 관측치 간의 거리로 계산한다(자세한 계산 방법은 2절 문헌 연구 참조). 해밍 거리는 직관적이고 이해하기 쉬우나, 범주형 데이터들이 가지고 있는 정보를 지나치게 단순화하는 단점을 가지고 있다. 해밍 거리의 단점을 Goodall(1966)과 Smironv(1968)은 확률방법(Probabilistic measure)에 기초하여 유사도 방법을 제안 하였고 Burnaby(1970)과 Lin(1988)은 정보이론(Information-theoretic measure)에 기초한 유사도 방법을 제안하였다.

Goodall(1966)은 두 변수 값의 유사도(두 변수 값이 동일한 경우 유사도는 가장 높음)를 측정 하는 방법으로 희소하게 발생하는(빈도가 낮은) 변수 값이 일치하는 경우(두 변수 값이 동일한 경우)에 상대적으로 더 높은 유사도를 부여하는 방식을 제안했다. 그러나 Goodall(1966)이 제안한 유사도 방법은 두 변수 값이 일치하지 않는 경우는 단순하게 0을 부여하였다. 이에 Smirnov(1968)는 두 변수 값이 일치하지 않는 경우에 0이 아닌 다른 비유사도 값을 부여하는 방법을 제시하였다. Smirnov(1968)는 두 변수 값이 일치하는 경우에는, Goodall(1966)과 유사하게, 빈도가 낮은 변수일 경우에 상대적으로 더 높은 유사도를 부여하는 방법을 채택하였으며, 두 변수 값이 불일치하는 경우에는 불일치하는 두 변수 값을 제외한 나머지 변수 값들을 고려하여 단순히 0이 아닌 낮은 유사도를 계산하는 방법을 제시하였다.

확률방법에 기초한 Goodall(1966)과 Smirnov(1968)의 방법은 두 변수 값이 일치하는 경우에 더 초점을 둔 유사도 계산 방법을 제시한 반면, Burnaby(1970)과 Lin (1988)은 정보이론에 기반을 두어 두 변수 값이 일치하지 않는 경우에 초점을 둔 방법이다. Burnaby(1970)는 두 변수 값이 일치하는 경우 단순하게 1의 유사도 값을 부여하였으며, 두 변수 값이 일치하지 않는 경우 희소한 두 변수 값의 불일치 값(즉, 비유사도 값)을 높게 부여함으로써 상대적

으로 유사도 값이 낮아지는 방법을 제시하였다. Lin (1988)은 Burnaby(1970)의 방법과는 다르게 빈번하게 발생하는 두 변수 값이 일치하는 경우 1이 아닌 높은 유사도 값을 부여하였으며, 희소하게 발생하는 두 변수 값이 불일치하는 경우 더 큰 비유사도 값을 부여하는 방법을 제시하였다.

그러나 앞서 언급한 모든 방법들은 변수와 변수 사이의 연관성을 고려하지 않는 단점을 지니고 있다. 연속형 변수에 대해서 마할라노비스(1936)는 연속형 변수 간 상관관계를 고려한 마할라노비스 거리(Mahalanobis distance)를 제시하였고, 이를 다양한 분야에서 활용하고 있으나, 범주형 자료에 대해서는 연관성을 고려한 거리 측정 방법에 관한 연구가 그리 흔치 않다. 이에 Le and Ho(2005)는 변수들 간의 연관성을 고려한 범주형 변수의 거리 측정 방법인 연관성 기반 비유사도(Association-based dissimilarity) 방법을 제안하였다. Le and Ho(2005)는 두 변수 값 사이의 비유사도를 다른 변수 값들의 조건부 확률 분포의 거리들을 이용하여 계산하는 방법을 제시하였다(자세한 계산 방법은 2장 문헌 연구 참조).

본 연구에서는 측정한 관측치가 범주형 변수로 구성되어 있을 때 변수간의 연관성을 고려한 관측치 간의 거리를 측정하여 군집분석을 수행하는 방법을 소개한다. 특히 변수 간 연관성을 고려한 방법(연관성 기반 비유사도)과 그렇지 않은 방법(해밍 거리)의 차이를 군집분석의 성능 비교를 통해 살펴보고자 한다. 이를 통해 만약 관측치를 구성하고 있는 범주형 변수들 간에 연관성이 존재하며 그러한 변수의 수가 증가할 경우, 연관성을 고려한 거리 측정 방법이 그렇지 않은 방법보다 더 나은 군집분석 성능을 보이는 것을 확인하고자 한다.

본 논문의 구성은 다음과 같다. 제2장에서는 연관성 기반 비유사도 방법과 PAM 군집분석의 적용에 관해 설명하고 제3장에서는 실험 설계 및 실험 결과에 관하여 기술하였다. 제4장에서는 이를 통한 결론을 서술하였다.

2. 제안방법: 연관성 기반 비유사성을 활용한 범주형 자료 군집분석

이번 장에서는 본 연구에서 제안하는 연관성 기반 비유사성을 활용한 범주형 자료 군집분석 방법을 설명하고자 한다. 본 제안방법은 크게 2단계로 구성된다; 1) 연관성 기반 비유사도 (Association-based dissimilarity)의 측정, 2) PAM (Partitioning Around Medoids) 군집분석의 적용. 설명의 편의를 위하여 본 장에서 사용되는 몇 가지 표기법에 대해서 정의하고자 한다. n개의 관측치로 이루어진 데이터 셋을 $\boldsymbol{X}=\{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$으로 표기하며, 각각의 관측된 관측치는 p차원의 범주형 변수로 구성되어있다. 범주형 변수 A_i는 q_i개의 변수 값을 가지며 $A_i = \{a_{i1}, a_{i2}, \cdots, a_{iq}\}$로 표기한다. 범주형 변수로 구성된 두 관측치의 거리는 먼저 범주형 변수의 변수 값들의 비유사도를 계산하고 이 비유사도를 이용하여 두 관측치 간의 거리를 계산한다. 본 논문에서 한 범주형 변수의 변수 값들의 비유사도는 $d(a_{ij}, a_{ik})$로 표기하며, 관측된 두 관측치의 거리는 $d(\boldsymbol{x_g}, \boldsymbol{x_h})$로 표기한다.

2.1 연관성 기반 비유사도 (Association-based dissimilarity)의 측정

먼저 첫 번째 단계에서는 범주형 변수들로 구성된 두 관측치 간의 거리를 측정하는 단계이다. 이를 위해서는 먼저 한 범주형 변수의 값이 다양하게 주어질 때 이들 값들 간의 비유사도를 측정하는 것이 필요하다. Le and Ho(2005)는 변수 값들의 비유사도를 한 변수의 변수 값이 주어졌을 때, 다른 변수의 변수 값들의 조건부 확률 분포의 차이로 계산하였다. Le and Ho(2005)의 연관성기반 비유사도를 계산하기 위해 특정한 변수 A_i의 변수 값 a_{ij}가 주어졌을 때 다른 변수 A_t의 변수 값 a_{tl}의 조건부 확률은 $p(a_{tl}|a_{ij})$로 표기하고, 보다 더 일반적으로 특정한 변수 A_i의 변수

값 a_{ij}가 주어졌을 때 다른 변수 A_t의 변수 값들의 조건부 확률 분포를 $\boldsymbol{P}(A_t|A_i=a_{ij})$로 표기하였다. 연관성 기반 비유사도 방법을 이용한 두 변수 값 사이의 비유사도는 아래와 같이 계산된다.

$$d(a_{ij},a_{ik})=\sum_{t=1}^{p}\psi(\boldsymbol{P}(A_t|A_i=a_{ij}),\boldsymbol{P}(A_t|A_i=a_{ik}))\ \ (t\neq i), \qquad \text{수식 (1)}$$

여기서 $\forall i,t\in\{1,2,\cdots,p\}$, $\forall j,k\in\{1,2,\cdots,q^i\}$이며, $\psi(\cdot,\cdot)$은 확률 분포의 거리 계산 함수이다. 확률 분포의 계산하는 방법에는 많은 연구자들에 의해 다양한 방법이 제안되어 왔다(Lin 1991; Kullback and Leibler 1951; Chakraborty 2008). 본 논문에서는 확률 분포 함수의 거리를 구하는 함수로 Hellinger 거리를 사용 하였다 (Chakraborty 2008).

Hellinger 거리는 두 확률 분포를 root vector로 표현했을 때 두 root vector의 차이의 Euclidean norm을 제곱근으로 나눈 결과와 동일하다. 즉, Hellinger 거리는 아래의 수식 (3)을 통해 계산된다.

$$\psi(\boldsymbol{P}(A_t|A_i=a_{ij}),\boldsymbol{P}(A_t|A_i=a_{ik}))=\sqrt{\sum_{l=1}^{q^t}(\sqrt{p(a_{tl}|a_{ij})}-\sqrt{p(a_{tl}|a_{ik})})^2} \qquad \text{수식 (2)}$$

여기서 $\forall i,t\in\{1,2,\cdots,p\}$, $\forall j,k\in\{1,2,\cdots,q^i\}$, $\forall l\in\{1,2,\cdots,q^t\}$이며, $p(\cdot|\cdot)$은 조건부 확률이다. Hellinger 거리는 비교하는 두 분포가 다를수록 Hellinger distance가 커지며, 두 분포가 완전하게 다를 경우 1을 가지며, 완전하게 동일할 경우 0을 가진다.

간단한 예제를 통해 Le and Ho(2005)의 연관성기반 비유사도의 계산법을 소개하고자 한다. 데이터 셋 $\boldsymbol{X}=\{\boldsymbol{x_1},\boldsymbol{x_2},\cdots,\boldsymbol{x_n}\}$가 색깔과 도형 두 범주형 변수로 구성되어 있으며(p=2), 색깔 변수는 하얀색(w)과 검은색(b)의 변수 값을 가질 수 있고, 도형 변수는 사각형(□), 다이아몬드(◇), 삼각형(△)의 변수 값을 가질 수 있다. 아래의 Table 1은 데이터 셋의 교차 빈도표와 교차 확률표를 나타낸다.

Table 1. The co-occurrence & conditional probability between Color and Shape

	Co-occurrence table			Conditional probability		
	white (w)	black (b)	sum	$p(w\|\cdot)$	$p(b\|\cdot)$	sum
□	40	30	70	4/7	3/7	1
◇	15	45	60	1/4	3/4	1
△	20	30	50	2/5	3/5	1

수식 (4)를 이용하여 (□,◇), (□,△), (◇.△)의 비유사도, 즉 $d(a_{ij},a_{ik})$를 구하면 아래와 같다.

$$d(□,◇)=\frac{1}{\sqrt{2}}\sqrt{(4/7-1/4)^2+(3/7-3/4)^2}=0.321,$$

$$d(□,△)=\frac{1}{\sqrt{2}}\sqrt{(4/7-2/5)^2+(3/7-3/5)^2}=0.171,$$

$$d(\diamond,\triangle)=\frac{1}{\sqrt{2}}\sqrt{(1/4-2/5)^2+(3/4-3/5)^2}=0.150.$$

다음으로 관측된 두 관측치 벡터 사이의 거리 $d(\boldsymbol{x_g}, \boldsymbol{x_h})$는 앞서 구한 비유사도의 합으로 아래와 같이 계산된다.

$$d(\boldsymbol{x_g}, \boldsymbol{x_h})=\sum_{i=1}^{p} d(a_{ij},a_{ik}), \qquad \text{수식 (3)}$$

여기서 $\forall\, g,h\in\{1,2,\cdots,n\}$이다.

2.2 PAM (Partitioning Around Medoids) 군집분석의 적용

두 번째 단계에서는 변수들 간의 연관성을 기반으로 측정된 두 관측치 간의 거리를 기반으로 군집분석을 수행하는 단계이다. 본 연구에서는 PAM (Partitioning Around Medoids) 군집분석을 활용한다. PAM 방법은 Kaufman과 Rousseeuw (1987)이 K-평균 군집방법을 수정 보완한 방법으로 K-평균 군집방법과 비슷하지만, K-평균 군집방법이 군집의 대푯값(cluster center)으로 동일 군집에 속한 관측치들 거리의 평균을 이용하는 반면, PAM 방법은 이상값과 결측값에 덜 민감한 medoid를 군집의 대푯값으로 사용한다. 여기서 medoid는 군집에 속한 관측치 중 하나를 의미한다. PAM 방법은 관측값과 가까운 medoid 간의 거리 합이 최소화되도록 하는 medoid를 구하는 방식으로 군집을 형성한다. 아래의 Table 2는 PAM 군집 분석의 알고리즘이다.

Table 2. PAM algorithm

PAM Algorithm
● Step 1: For a given cluster assignment C find the observation in the cluster minimizing total distance to other points in that cluster: • $i_k^* = \arg\min_{\{i:\, C(i)=k\}} \sum_{C(i')} D(x_i,x_{i'})$. • Then $m_k = x_{i_k^*}$, $k=1,2,\cdots,K$ are the current estimates of the clusters.
● Step 2: Given a current set of cluster centers (a.k.a. medoids) $\{m_1, \cdots, m_K\}$, minimize the total error by assigning each observation to the closest (current) cluster center: • $C(i) = \arg\min_{\{1\le k\le K\}} D(x_i,m_k)$.
● Step 3: Iterate steps 1 and 2 until the assignments do not change

PAM 방법 특징으로는 관측치의 좌표 값이 아닌 각 관측치들의 거리를 계산한 거리행렬을 이용해도 군집분석이 가능하다는 점에 있다. 본 연구에서 다루는 범주형 변수는 수치로 표시된 좌표로 나타낼 수 없으며, 다만 각 관측치들 간의 거리만을 계산할 수 있다. 즉 연관성 기반 비유사도 방법은 각 관측치 사이의 거리를 계산하기 때문에 본 연구에서는 PAM 방법을 군집분석의 방법으로 선정하였다.

3. 성능평가 실험 설계 및 결과

본 3장에서는 본 연구에서 제안하는 연관성 기반 비유사도 측정방식과 연관성을 고려하지 않는 거리 측정방식의 예로서 가장 대표적인 해밍 거리(Hamming distance)를 이용한 방법을 각각 활용하여 군집분석을 수행하고 그 결과를 비교한다. 특히 연관성이 있는 범주형 변수의 수가 증가함에 따라 두 다른 거리 측정 방법이 군집분석의 성능에 미치는 영향을 살펴보고자 한다. 본 연구에서 군집분석 방법은 제안방법에서 사용했던 PAM (Partitioning around medoids) 군집분석 방법(Kaufman and Rousseeuw 1987)를 동일하게 활용하고, 군집분석 성능 척도로는 Rand Index (Rand 1971)를 사용하였다.

3.1 비교대상: 해밍 거리 (Hamming distance)

해밍 거리는 리처드 해밍이 1950년에 해밍 부호와 함께 도입한 거리 함수로 범주형 변수의 거리를 계산하는 가장 직관적이고 쉬운 방법이다. 해밍거리에서 변수 값들의 비유사도는 아래와 같이 계산된다.

$$d(a_{ij}, a_{ik}) = \begin{cases} 0 & \text{if } a_{ij} = a_{ij} \\ 1 & \text{if } a_{ij} \neq a_{ij} \end{cases}, \qquad \text{수식 (4)}$$

여기서 $\forall i \in \{1,2,\cdots,p\}$이며, $\forall j,k \in \{1,2,\cdots,q^i\}$ 이다. 즉, 해밍거리에서 두 변수 값의 비유사도는 동일한 값을 가지면 0이고 다른 값을 가지면 1이다.

3.2 데이터 셋

본 실험에서 활용된 데이터는 UCI repository에서 범주형 자료들의 상관관계가 존재하는 데이터 셋 4개(Credit card, Breast cancer, Lymphography, Mushroom)를 선정하였다 (http://archive.ics.uci.edu). Credit card는 신용카드 승인에 관한 데이터로 총 9개의 범주형 변수와 6개의 연속형 변수로 구성되었다. 본 연구의 초점은 범주형 변수와 연속형 변수를 모두 다루는 방법이 아니라 범주형 변수를 다루는 방법에 한정하고 있으므로 본 연구에서는 범주형 변수로만 분석을 진행하였다. Credit card의 범주형 변수는 민감한 정보를 가지고 있어 모두 의미 없는 기호들로 대체되어있다. 다음으로 Breast cacner 데이터는 유방암의 진단에 관한 데이터로 총 9개의 범주형 변수로 구성되어있다, Lyphography는 림프관 조영 촬영 방법에 대한 진단 데이터로 총 18개의 변수로 구성이 되어있으나 이 중 3개의 변수는 연속형 변수를 이산화(Discretize)한 변수여서 3개 변수를 제외한 15개의 범주형 변수만을 분석에 사용하였다. 마지막으로 Mushroom 데이터는 식용버섯과 독버섯의 구분에 필요한 정보를 측정한 데이터이다. 총 22개의 범주형 변수로 구성되있다. 각 데이터에 대한 관측치의 수, 범주형 변수의 수, 알려진 군집의 수는 아래의 Table 3와 같다. Lyphography 데이터를 제외한 나머지 3개의 데이터는 모두 결측값를 가지고 있다. 결측값를 다루는 여러 방법이 존재하지만, 본 연구에서는 결측값를 가지고 있는 관측치를 제거하는 간단한 방법을 이용하였다.

Table 3. Summary of the data set

	Credit card	Breast cancer	Lymphography	Mushroom
Number of observations(n)	653	683	148	5,644
Number of variables(p)	9	9	15	22
Number of clusters	2	2	4	2

3.3 군집분석 성능평가 척도

본 연구에서는 군집분석의 성능을 비교하는 척도로 Rand(1971)가 제시한 Rand Index를 사용하였다. Rand Index는 두 군집분석의 결과가 서로 유사한지 비교하는 방법으로, Rand Index의 분자는 동일한 두 관측치 쌍이 동일한 군집에 소속되어 있는가, 혹은 상이한 군집에 소속되어 있는가에 대한 빈도수이며, 분모는 전체 관측치 쌍의 수이다. Rand Index는 Table 4와 같은 2×2 분할표를 사용하여 구할 수 있다. Table 4에서 짝지어진 관측치 쌍의 모든 경우의 수는 n(n-1)/2개이며, 이들은 다음의 두 범주로 분류된다.

- 1: 짝지어진 쌍이 군집분석 결과에서 동일한 하나의 군집에 속한 경우
- 0: 짝지어진 쌍이 군집분석 결과에서 서로 다른 군집에 속하는 경우

Table 4. 2 x 2 Contingency matrix

		Clustering result B	
Clustering result A		1	0
	1	a	b
	0	c	d

여기서 a는 군집분석 결과 A와 B 모두에서 두 쌍의 관측치가 동일한 군집에 속하는 경우의 수를 의미하며, b는 군집분석 결과 A에서 두 쌍의 관측치는 동일한 군집에 속하지만, 군집분석 결과 B에서 두 쌍의 관측치가 서로 다른 군집에 속하는 경우의 수를 의미한다. c는 군집분석 결과 A에서 두 쌍의 관측치는 서로 다른 군집에 속하지만, 군집분석 결과 B에서는 서로 동일한 군집에 속하는 경우의 수를 의미하며, 마지막으로 d는 군집분석 결과 A와 B 모두에서 두 쌍의 관측치가 서로 다른 군집에 속하는 경우의 수를 의미한다. 따라서 a와 d의 경우는 두 군집분석의 결과가 서로 동일함을 의미하고 b와 c는 서로 상이함을 의미한다. RAND Index는 총 개체 쌍 $(a+b+c+d)=\binom{n}{2}$ 중 일치하는 쌍의 비율이며 아래와 같이 계산된다.

$$RI=\frac{a+d}{a+b+c+d}$$

Rand Index 값이 1이면 두 군집분석의 결과는 완전히 동일함을 의미하고 Rand Index의 결과 값이 0 이면 두 군집분석의 결과는 완전히 상이함을 의미한다. 본 연구에서는 군집에 대한 정보가 이미 존재하여, 실제 군집에 대한 정보와 PAM 군집분석 결과와 비교함으로써 군집분석의 성능을 평가하였다. 즉 Rand Index 값이 1이 나오는 것은 PAM 군집분석을 통해 나온 군집과 실제 군집의 결과가 완전히 동일함을 의미하며, PAM 군집분석을 통해 실제 군집을 정확하게 복원할 수 있음을 의미한다.

3.4 실험 결과

본 실험에서는 변수의 수에 따라 군집분석 성능의 변화를 살펴보기 위하여 임의로 변수를 선택하여 표본을 구성하고 표본을 이용한 PAM 군집분석의 Rand Index 값을 비교 분석하였다. Credit card 데이터와 Breast cancer 데이터의 경우 총 9개의 변수로 구성되어 있으며, 각각 3개, 5개, 7개의 변수를 임의로 선택하여 표본을 구성하고 Rand Index 값을 계산하였다. Lymphography 데이터의 경우 총 15개의 변수로 구성되어 있으며, 5개, 10개, 13개의 변수를 임의로 선택하여 표본을 구성하였고, 마지막으로 Mushroom 데이터는 총 22개의 변수로 구성되어 있으며, 각각 10개, 15개, 20개의 변수로 표본을 구성하여 Rand Index 값을 계산하였다. 각 표본을 구성하는 변수를 선택하는 과정은 임의로 선택하였으며 100회 반복하여 표본을 생성하였다.

Table 5. Comparison of Rand Index on four real-life data set

		Hamming distance		Association-based dissimilarity		
		Rand Index		Rand Index		
Dataset	# of variables	Mean	Standard deviation	Mean	Standard deviation	Difference
Credit card	3	56.43 %	9.04 %	**58.12 %**	10.30 %	1.69 %
	5	55.22 %	6.35 %	**60.09 %**	10.13 %	**4.87 %*****
	7	54.42 %	6.45 %	**59.07 %**	9.53 %	**4.65 %*****
Breast cancer	3	66.67 %	9.84 %	**86.43 %**	4.45 %	**19.76 %*****
	5	74.70 %	8.65 %	**89.18 %**	2.36 %	**14.48 %*****
	7	79.79 %	2.32 %	**90.37 %**	1.37 %	**10.58 %*****
Lymphography	5	55.00 %	2.54 %	55.00 %	2.52 %	0.00 %
	10	55.89 %	2.16 %	56.51 %	2.36 %	0.62 %*
	13	56.19 %	1.44 %	**57.42 %**	2.02 %	**1.23 %*****
Mushroom	10	70.09 %	6.94 %	**73.64 %**	5.74 %	**3.55 %*****
	15	72.39 %	3.82 %	**74.88 %**	0.90 %	**2.49 %*****
	20	74.30 %	0.80 %	**74.91 %**	0.16 %	**0.61 %*****

Note * p < 0.1, **p < 0.05, ***p< 0.01

Table 5는 각 분석 결과에 대한 평균과 표준편차를 요약한 결과이며, Figure 1은 각 표본에 대한 표본 평균과 95% 신뢰구간을 의미한다. 먼저 Credit data를 살펴보면 해밍거리를 이용한 PAM 군집분석(이후 HB_clustering;

Hamming distance Based clustering)의 결과(56.43%, 55.22%, 54.42%)보다 연관성기반 비유사도를 이용한 PAM 군집분석(이후 AB_clustering; Association-based dissimilarity Based clustering)의 결과(58.12%, 60.09%, 59.07%)가 더 높게 나타났다. 두 결과의 차이를 통계적으로 검증해본 결과 표본을 구성하는 범주형 변수가 3개인 경우에서는 두 결과 차이(1.69%)는 통계적으로 유의하지 않게 나타났다. 그러나 범주형 변수가 5개와 7개로 구성된 경우에서는 그 차이(4.87%와 4.65%)가 모두 유의수준 0.01에서 통계적으로 유의하게 나타났다.

다음으로 Breast Cancer에서의 HB_clustering의 Rand Index는 각각 66.67%, 74.70%, 79.79%로 나타났으며, AB_clustering의 Rand Index는 각각 86.43%, 89.18%, 90.37%로 나타났다. 통계적 결과 모든 평균의 차이(19.76%, 14.48%, 10.58%)는 유의수준 0.01에서 통계적으로 유의하게 나타났다.

Lymphography의 경우 HB_clustering의 Rand Index는 각각 55.00%, 55.89%, 56.19%로 나타났으며, AB_clustering의 Rand Index는 각각 55.00%, 56.51%, 57.42%로 나타났다. 표본을 구성하는 범주형 변수가 5개인 경우 두 평균의 차이는 나타나지 않았으며, 범주형 변수가 10개로 구성된 표본에서의 평균의 차이는 0.62%이다. 이 차이에 대한 통계적 검증결과 유의수준 0.1에서 유의하게 나타났다. 마지막으로 범주형 변수가 13개로 구성된 표본에서의 평균의 차이는 1.23%이며, 차이에 대한 통계적 검증결과 유의수준 0.01에서 유의하게 나타났다.

마지막으로 Mushroom 데이터를 살펴보면 HB_clustering의 Rand Index(70.09%, 72.39%, 74.30%)보다 AB_clustering의 Rand Index(73.64%, 74.88%, 74.91%)가 높게 나타났다. 각각의 차이에 대한 통계적 검증결과 모든 차이(3.55%, 2.49%, 0.61%)에 대해 유의수준 0.01에서 통계적으로 유의하게 나타났다.

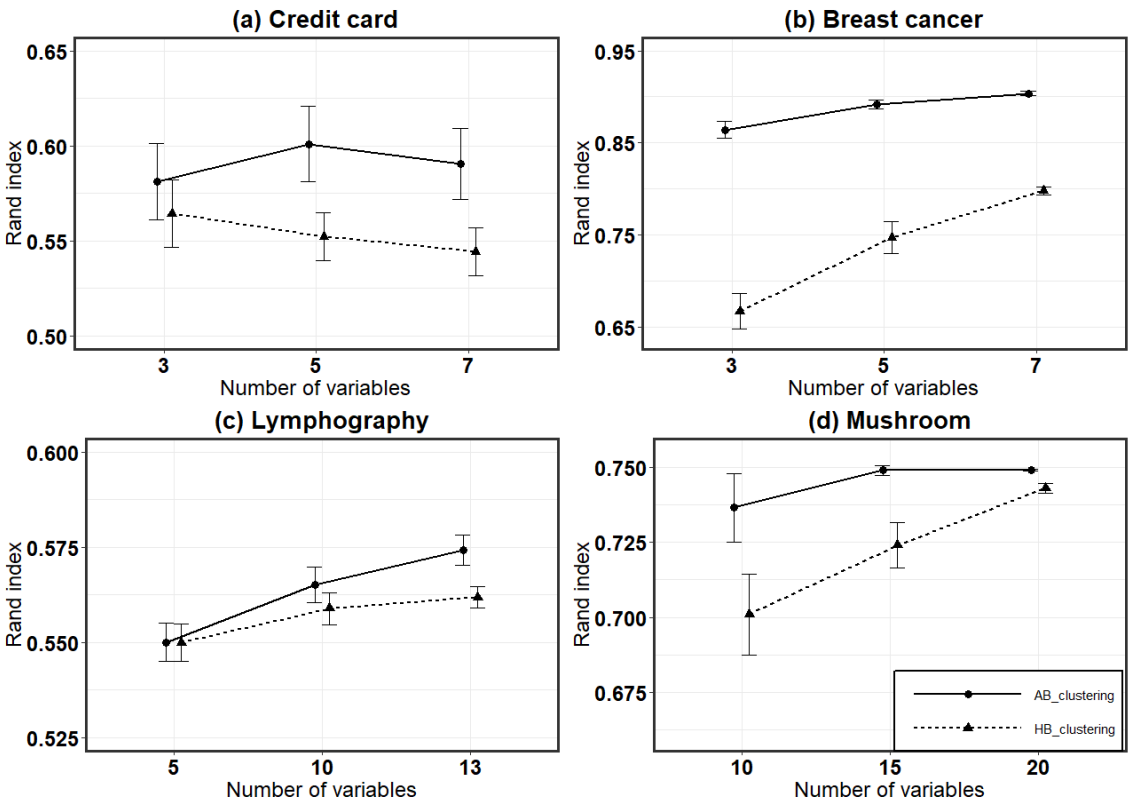

Figure 1. Rand index comparison results

본 실험 결과를 살펴보면 다음과 같은 결론을 도출할 수 있다.

범주형 자료에 대한 군집분석 등과 같은 데이터마이닝 방법을 적용함에 있어 연속형 변수를 다룰때와 유사하게 연관성을 고려할 필요가 있다. 본 연구의 실험을 결과를 보면 대부분의 경우에서 연관성 기반 비유사도를 이용한 군집분석의 성능이 해밍 거리를 이용한 군집분석의 성능보다 높게 나타났다. 각 표본 평균의 통계적 차이를 검증한 결과 2가지 경우(Credit card에서 3개의 변수로 표본을 구성한 경우와 Lymphography의 5개의 변수로 표본을 구성)를 제외하고 모두 통계적으로 유의한 차이를 보였다. 또한, 관측치를 구성하고 있는 변수의 수가 증가함에 따라 연관성 기반 비유사도를 이용한 군집분석 성능과 해밍 거리를 이용한 군집분석의 성능 차이가 통계적으로 유의하게 나타났다. Credit card에서 3개의 변수로 표본을 구성한 경우와 Lymphography의 5개의 변수로 표본을 구성한 경우를 보면 두 군집분석의 성능의 차이가 통계적으로 유의하지 않게 나타났으나, 표본을 구성하는 변수를 증가시켜 감에 따라 두 군집분석의 성능의 차이가 통계적으로 유의하게 나타났다. 이 결과를 토대로 볼 때 범주형 변수로 구성된 데이터의 경우에도 연속형 변수로 구성된 데이터와 같이 연관성을 고려하여 군집분석과 같은 데이터마이닝 방법을 적용하는 것이 중요함을 의미한다고 볼 수 있다.

다음으로 Credit card 데이터를 제외하곤 변수의 수가 증가함에 따라 군집분석의 성능도 향상하지만, 그 향상 폭은 감소하는 현상을 보인다. 이는 군집분석에 사용되는 변수의 수를 줄임으로써 군집분석의 성능은 다소 감소시킬 수 있으나, 계산 시간 및 저장 공간 등의 분석 비용이 절감하는 효과를 거둘 수 있음을 의미한다. 즉 범주형 자료 역시 변수선택(Feature selection)의 문제를 고려할 필요가 있다고 볼 수 있다. 범주형 자료로 구성된 데이터셋에서의 변수선택 문제는 본 연구의 범위를 벗어남으로 이는 후속 연구를 통해 보완하고자 한다.

4. 결론

연속형 변수의 경우 상관관계와 같은 데이터의 특징을 반영하는 다양한 거리 측정 방법이 논의됐으며 다양한 군집분석 모형과 분류 모형에 폭넓게 활용됐으나 범주형 변수의 경우 데이터의 특징을 반영하는 거리 측정 방법에 대한 논의는 상대적으로 적은 관심을 받아왔으며 이를 군집분석과 분류 모형에 활용하는 방안에 대해서도 상대적으로 더디게 진행됐다. 이에 본 논문은 범주형 변수의 연관성을 고려한 연관성 기반 비유사도 방법을 소개하고 이를 활용한 군집분석의 방법론을 소개하였다. 그리고 본 논문에서 제시한 방법의 효과성을 검증하기 위해 해밍 거리를 이용한 군집분석의 성능과 연관성 기반 비유사도를 이용한 군집분석의 성능을 비교하였다. 실험 결과에 따르면 범주형 변수의 연관성이 존재하는 경우 임의로 범주형 변수를 선택하여 표본을 구성하였음에도, 이를 무시하는 거리 측정 방법보다 연관성을 고려하는 거리 측정 방법이 더 우수한 군집분석 성능을 보였다.

본 연구의 시사점으로는 범주형 자료의 수집이 잦은 금융업, 의료업과 같은 분야의 기업들이 데이터를 분석함에 있어 본 연구에서 소개한 방법을 통해 더욱 성능이 우수한 군집분석 결과를 누릴 수 있을 것이라는 시사점을 제공한다. 본 연구의 뒤를 이을 향후 연구주제로는 변수선택(Feature selection)방법을 이용하여 군집분석 성능 향상에 이바지하는 핵심 변수를 찾는 방법과 범주형, 연속형 변수들이 혼재해 있는 경우에서의 거리측정 방법을 개발하는 것이다. 이러한 후속 연구들은 품질경영 영역에서 범주형 자료를 보다 효과적으로 활용한 흥미로운 주제가 될 것으로 보인다.

REFERENCES

Burnaby, T. P. 1970. "On a method for character weighting a similarity coefficient, employing the concept of information." Journal of the International Association for Mathematical Geology 2(1):25-38.

Cha, S. H. 2007. "Comprehensive survey on distance/similarity measures between probability density functions." City 1(2):1.

Chakraborty, D. D. 2008. Statistical decision theory. estimation, testing and selection. Investigación Operacional 29(2):184-185.

Esposito, F., Malerba, D., Tamma, V., & Bock, H. H. 2000. "Classical resemblance measures. Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data," 15, 139-152.

Goodall, D. W. 1966. "A new similarity index based on probability." Biometrics, 882-907.

Hamming, R. W. 1950. "Error detecting and error correcting codes." Bell System technical journal 29(2):147-160.

Jia, H., Cheung, Y. M., & Liu, J. 2016. "A new distance metric for unsupervised learning of categorical data." IEEE transactions on neural networks and learning systems 27(5):1065-1079.

Kaufman, L., & Rousseeuw, P. 1987. Clustering by means of medoids. North-Holland.

Kullback, S., & Leibler, R. A. 1951. "On information and sufficiency." The annals of mathematical statistics 22(1):79-86.

Le, S. Q., & Ho, T. B. 2005. "An association-based dissimilarity measure for categorical data." Pattern Recognition Letters 26(16):2549-2557.

Lim, Y. B., Kim, S. I., Lee, S. B., & Jang, D. H. 2016. "Literature Review on the Statistical Methods in KSQM for 50 Years." Journal of the Korean Society for Quality Management 44(2):221-244.

Lin, D. 1998. "An information-theoretic definition of similarity." In Icml 98(1998), 296-304.

Lin, J. 1991. "Divergence measures based on the Shannon entropy." IEEE Transactions on Information theory, 37(1):145-151.

Mahalanobis, P. C. 1936. On the generalized distance in statistics. National Institute of Science of India.

Rand, W. M. 1971. "Objective criteria for the evaluation of clustering methods." Journal of the American Statistical association 66(336):846-850.

Seo, M. K., & Yun, W. Y. 2017. "Clustering-based Monitoring and Fault detection in Hot Strip Roughing Mill." Journal of the Korean Society for Quality Management 45(1):25-38.

Smirnov, E. S. 1968. "On exact methods in systematics." Systematic Biology 17(1):1-13.

Suh, C. J., Kim, H.T., Kim, J.H., Kawk, Y.W.. 2013. Introduction to Management Quality: 1st edition: Parkyong.